

V. Data Control and Preparation

This chapter describes the procedures used to transform responses from the student questionnaire into a computerized data file. These procedures include editing completed questionnaires for missing information, retrieving the missing information, monitoring the receipt of completed questionnaires, preparing the questionnaires for data entry, and preparing the documents for archival storage. To efficiently accommodate the large number of respondents and the many variables constituting the NELS:88 student survey, most of the questions in the student questionnaire and eighth grade tests used response formats suitable for optical mark reading.

5.1 Onsite Editing and Retrieval

The first part of the data control process involved editing questionnaires and retrieving missing information. NORC field staff conducted onsite editing of the student questionnaires by first checking that the student identification number was correctly filled in. Next, the "critical items," so designated because of their special interest to analysts, their policy relevance, or their usefulness in locating the student for subsequent follow-up studies, were checked for completeness. (A complete listing of the critical items appears in Appendix B.)

If the response to one or more of the critical items was missing, undecipherable, or had multiple categories marked when only one response was required, the NORC field staff member privately pointed out the problem to the student. If, after prompting, the student indicated that he or she had chosen not to answer the question, the NORC staff member marked a "no retrieval" response for the item. (No retrieval was indicated by filling in an oval positioned to the left of each critical item). The "no retrieval" responses were used later during the machine editing process to assign a "refused" response to the critical items. Most editing and retrieval for the student questionnaire was conducted in the way just described. In a very small number of instances (fewer than 300 cases), additional retrieval of missing responses to critical items had to be conducted after the questionnaire reached NORC.

A small number of student questionnaires were administered by school coordinators rather than NORC personnel and were not subject to onsite editing and retrieval (see section 4.3.1). These cases reflect small schools with only one or two eligible eighth graders or make-up sessions with fewer than five students. To ensure respondent confidentiality, these questionnaires were reviewed for completeness by authorized personnel upon receipt at NORC. The editing process involved a review of all critical items and the retrieval of missing (and/or inappropriately marked) items by experienced NORC telephone interviewers. Student responses were recorded in the questionnaire. The "retrieval oval" was marked to indicate that an attempt was made to retrieve the item and the appropriate response category was filled in accordingly. The retrieval was begun in early June, 1988 and completed in late August, 1988.

5.2 Monitoring and Receipt Control

After completing data collection and onsite editing, NORC field staff prepared the survey materials and tests for mailing to NORC. Once these packages were received at NORC they passed through several steps. First, receipt control clerks checked each student questionnaire for completeness and reviewed the transmittal documents to ensure that the case ID numbers matched. A final disposition code was assigned to the corresponding student by the team leader. The disposition code indi-

cated whether test data, questionnaire data, or both were completed by that student. Receipt control clerks then entered this disposition code into NORC's Survey Management System (SMS), a micro-computer-based system that replaced the NORC Automated Survey System (NASS) used on earlier studies. At the time of entry, the SMS generated and automatically entered the date that data for each case was received.

5.3 Inhouse Editing and Coding

The next step was to edit and code the confidential locator pages from the questionnaire and to separate these pages from the rest of the questionnaire. A section of the student questionnaire asked students to provide identifying information and information about their parents' occupations. This handwritten locating information was edited for legibility. NORC coders used a coding procedure to condense the occupation questions into the eighteen categories used in the occupation questions in the parent questionnaire. (A list of the occupation categories can be found on page 14 of the parent questionnaire in question 34B.) In coding this item it was discovered that eighth graders classified one or the other of their parents as a "student" a sufficient number of times to justify the creation of a nineteenth category.

5.4 Data Entry and Archival Storage

When editing, coding, and inhouse retrieval were completed, questionnaires were separated into two parts, each of which received different treatment with respect to data entry and archiving. First, the seven pages of identifying information were removed from each questionnaire. This information was sent to NORC's data entry department for processing. When data entry was completed, the pages were filed such that they would be accessible for use during the remaining phases of the survey.

The data entry for the remaining part of the each questionnaire, which contained students' responses to the majority of the questions, was completed through an optical mark reading procedure. Optical mark reading was conducted by NORC's subcontractor, Questar Data Systems, Inc., which received the questionnaires and tests in batches for processing. Questar also arranged to have questionnaires and tests photographed onto microfilm. Once the questionnaires were scanned and photographed they were destroyed and the rolls of microfilmed questionnaires were returned to NORC for archival storage.

VI. Data Processing

Data processing activities span the entire length of the NELS:88 base year student survey, beginning with drawing the sample, continuing with receipt control and machine editing, and ending with the preparation of public use data tapes and user documentation.

6.1 Student Locator Database

The locator database contains the most up-to-date name and address information available for each student. These data were constructed both from the sample file and from locating information provided by the student, and so contain the data required to trace a student through the school or district. Locating information was provided in Part I of the student questionnaire, including the student's name and address, his or her parents' names and address(es), and the name, address, and relationship of another person likely to stay informed of the respondent's whereabouts. To ensure confidentiality, all identifying information is stored on secure files that are separate from the questionnaire data. Part I of the student questionnaire also requested information regarding respondent birth date, sex, parent occupation, and the sector (e.g., public, private) of the high school he or she expected to attend. These data are included in the public use data tapes.

6.2 Receipt Control Procedures

The NORC Survey Management System (SMS) was used to track survey activities. This system houses a record for each student that contains the school ID, the respondent ID number, student and parent disposition codes, and other information. Data control disposition codes in the SMS files were used to track completion rates of the sample during the data collection. At the end of the data collection period the SMS file was merged with the scanned data to search for any discrepancies in IDs or final status. In most cases, it was possible to resolve such discrepancies by referring to the microfilm of the documents.

6.2.1 Storage and Protection of Completed Instruments and Records

Whenever questionnaires were not being processed, they were filed in locked cabinets. After data retrieval and editing, the locator pages containing the respondent's name and ID were data-entered into the student locator database, then detached and filed in a locked cabinet, in a locked room. From this point on, the respondent's name and address could no longer be associated with his or her responses to the questionnaire. Questionnaires were stored in locked file cabinets in locked rooms until they were transmitted to the scanning subcontractor, who observed identical security and confidentiality protection safeguards. The optical scanning subcontractor for the NELS:88 base year was Questar Data Systems, Inc.

6.3 Optical Scanning

With the exception of the student locator section, NORC used the optical mark read (OMR) method of data conversion for the student questionnaire and eighth grade tests. (Key-to-disk equipment at NORC was used to convert the locator section to machine readable form.) The materials were optically scanned using equipment that read darkened ovals or marks on the page. The scanning subcontractor conducted extensive tests and checks of the machine's ability to correctly read the darkened ovals. Adjustments were made to the marksense threshold as required. To check the accuracy of data conversion, the scanning programs were tested in two ways: through use of dummy question-

naires specifically designed to detect scanning errors, and by running a substantial number of real documents through the system. Final data were compared item by item to hard-copy questionnaires, and procedures were modified until accuracy was attained.

6.4 Machine Editing

Conventions for editing, coding, error resolution, and documentation adhered as closely as possible to the procedures and standards previously established for HS&B and NLS-72.

After the scanning contractor completed data conversion and supplied NORC with a raw data tape, the combination of machine editing and visual inspection of the output began. The tasks performed included: resolving inconsistencies between filter and dependent questions, supplying the appropriate missing data codes for questions left blank, and detecting illegal codes and converting them to missing data codes. Variable frequencies were inspected before and after these steps to verify the correctness of the automated processes.

Inconsistencies between filter and dependent questions were resolved in the machine editing process. In most instances, dependent questions that conflicted with the skip instructions of a filter question contained data that, although possibly valid, were superfluous. For instance, respondents sometimes indicated "no" to the filter item and then continued to answer "no" to subsequent dependent questions. If a value was given to a filter question indicating that the respondent should have skipped the subsequent question(s), those questions were set to a value of legitimate skip even if the respondent answered some or all of these questions. If a multiple response or no answer was given to a filter question that was not a legitimate skip, the question was assigned an appropriate reserve code ("6", "7", or "8") and all subsequent questions that might have been skipped were processed as if the respondent should have answered them.

After improperly answered questions were converted to blanks, the student data were passed through a second step in the editing program that supplied the appropriate reserve codes for blank questions. Where a value was not provided by the respondent, a reserve code fills the field. These codes are as follows:

6 = MULTIPLE RESPONSE

7 = REFUSED (if a critical item is missing and the retrieval oval is checked)

8 = MISSING

9 = LEGITIMATE SKIP

If the field is longer than one column, the right-hand column contains one of the above codes and the rest of the columns are filled with "9"s.

Each critical item has an associated "retrieval oval." The retrieval oval was marked if an attempt was made to retrieve data from a respondent. These flags then were used to set corresponding blank data to REFUSED. Although retrieval variables were present in the questionnaire, they are not present in the data since their purpose was to determine correct reserve codes. Any critical item that was blank, not a legitimate skip, and whose respective retrieval oval flag was checked was coded as "7" (refused). A critical item that was blank, not a legitimate skip, and whose respective retrieval flag was not checked was coded as "8" (missing). If a filter was coded "7" (refused), all subsequent questions that might have been skipped were processed as if the respondent should have answered them. Filters that were coded "6" (multiple response) or "8" (missing) were handled the same way.

Detection of out-of-range codes was completed during scanning for all questions except those permitting an open-ended response. The two-digit occupation codes for the manually coded, open-ended questions were checked manually to validate all codes.

The frequency with which responses were recoded to legitimate skip for each skip pattern was closely monitored. Frequency distributions of responses before and after editing were inspected. All filter questions and their respective dependent items were displayed in condensed crosstabulations so that staff could verify the correctness of the recoding.

6.5 Data File Preparation

The conventions used to assign SAS and SPSS variable names are as consistent as possible with HS&B and NLS-72. In those two surveys, variable names were assigned according to the survey wave and the question number. A similar system was developed for NELS:88. For example, BYS56A, is from the base year student survey, question 56, part A.

Most composite variables were constructed using responses from two or more questionnaire items. In some cases, composites were constructed from variables from different databases. Others were constructed by recoding a variable and a very few were simply copied from a different data source to this file for the user's convenience. Composite variables may be valid throughout the survey (e.g., SEX) or they may be specific to this particular survey wave. The names of the latter begin with BY for base year. Hence, BYFAMSIZ categorizes the base year family size. Weights are similarly labeled: BYQWT for the selection weight for questionnaire completion adjusted for nonresponse during the base year, and so on. Composite variables, such as SEX, RACE, or G8ENROL, which will remain the same throughout the survey waves, have names that will remain the same.

The only reserve code used for composite variables is that of missing data. For one-column variables that is an 8, for variables greater than one column, the left most columns are filled with "9"s (9....8). This reserve code is used when the sources for data are either item nonresponse or nonparticipation in all or part of the components of the study. Appendix D contains explanations of the conditions under which specific composite variables were assigned a missing code.

VII. Guide to the Data Files and Codebook

The NELS:88 public use data files are available on four separate tapes, one for each study component: the student survey, the parent survey, the teacher survey, and the school administrator survey. The tape for the student survey component contains a data file based on data for 24,599 participating students from 1,052 schools, including the OBEMLA student oversamples. As indicated earlier, the student data can be used alone or merged with the parent, teacher, or school files.

Since multiple instruments were used to gather data from students, parents, teachers, and school administrators, the analyst must use the proper participation flags and weights to produce accurate statistics. Therefore, before describing the data files, several suggestions are offered that should be helpful to the analyst. These are followed by a complete description of the content and organization of the student data file and a guide to the associated codebook.

In the section on the data file, the reader should pay particular attention to the composite variables, which have been especially constructed to streamline substantive analyses. Since researchers often need to control for education level, urbanicity of school, educational aspirations, socioeconomic status, and the like, a set of classification variables has been carefully constructed that can be used for this purpose. Complete specifications used to create these composite variables can be found in Appendix D. Should the analyst choose to create alternatives, he or she is, of course, free to do so.

7.1 Suggestions for Selecting Participation Flags and Weight and Using Statistical Programs

One of the first steps to take before running statistical analyses is to select the proper participation flags and weight. There are seven participation flags (BY indicates base year) which define subsets of the participating students (those who completed the student questionnaire).

For the following six flags, a 1 specifies that the indicated documents (questionnaires and/or tests) were completed, and a 0 that they were not.

BYTEQFLG	at least one teacher questionnaire
BYPAQFLG	a parent questionnaire
BYTXPAFG	the student tests and a parent questionnaire
BYTEPAFG	a parent questionnaire and at least one teacher questionnaire
BYTXFLG	the student tests
BYADMFLG	the school administrator questionnaire

BYIEPFLG, the seventh flag, is 1 if the student had on file an Individualized Education Program and was reported to the Department of Education as belonging to one of the following handicap categories: deaf, hard of hearing, deaf-blind, or multiple handicap (only if hard of hearing was included as one of his or her impairments); and the student is currently mainstreamed with regular hearing eighth grade students for English or mathematics classes. It is 0 if the above criteria were not satisfied.

These flags should be used to select the subset of respondents the analyst intends to examine. For example, if data are desired from all students for whom a parent questionnaire and at least one teacher questionnaire was completed, BYTEPAFG should be used to select them. (Even when running unweighted statistics, the participation flags should be used). When the user combines a flag with the

appropriate weight, he or she can produce population estimates. There are two weights for NELS:88 data: BYQWT, included on the student file for producing weighted student statistics; and BYADMWT on the school file for producing weighted statistics for schools with participating students.

To compute a weighted estimate of the proportion of students, with corresponding teacher data, who felt that the teachers in their school were interested in students (question 59G), for example, one would take the following steps:

- (1) use the participation flag, BYTEQFLG, to select the cases for whom a teacher questionnaire was completed;
- (2) invoke the appropriate weight, BYQWT; and
- (3) run frequencies for the variable BYS59G.

The appropriate participation flags and/or weights should be used if unweighted and weighted analyses are to be performed correctly. See Appendix F for specific examples using Statistical Analysis System (SAS).

Although sampling weights are discussed in detail in Chapter III, a few words are warranted here. The NELS:88 data files are designed to be used as weighted data sets in all analyses. The complexity of the sample design of the base year virtually ensures inaccurate results if the data are analyzed on an unweighted basis. Clustering, multistage selection, and disproportionate sampling all contribute potential bias and various degrees of unreliability, which can be avoided by using the weights provided to analyze specific subsets of the sample.

7.1.1 Packaged Statistical Programs

NCES has responded to numerous questions over the years having to do with statistical analyses of data from earlier longitudinal education studies and now routinely recommends the procedures outlined in Appendix F, using SAS with NELS:88 data. SPSS-X can also be used, and the data tape contains the appropriate control cards for this package. Analysts should contact their own support facilities to obtain the information necessary to create an SPSS-X system file from a SAS system file and vice versa.

7.2 Content and Organization of the Data Files

The student raw data file consists of 24,599 records for participating students. (Nonparticipating students are not included on the base year data tape of a longitudinal study). Each record is organized as shown in the record layout that appears as Appendix C. The variables on the record are grouped into logical sets as discussed below. For the sake of brevity, each item of data is referred to by its SAS (SPSS-X) variable name, as defined in the control cards provided with the data file.

The student data tape contains four related files. They are:

1. The raw data file, with items in the following order for each respondent:
 - a. Randomized ID number (positions 1-7)
 - b. Information from the student questionnaire (positions 8-358)
 - c. Base year weight, flags, and composites (positions 359-577)

2. SPSS-X control cards
3. SAS control cards
4. SAS system file

Questionnaire data from school administrators, students, or both sources were collected from 1,057 schools in the core sample. Five of these 1,057 schools were dropped from both the school and the student data files because student questionnaire data were missing, leaving 1,052 schools either with school administrator and student data, or with student data only. These 1,052 schools are represented on the student file.

For 17 of the 1,052 schools no school administrator data were available. Because these 17 schools are not included in the school file (which contains as its main source of data responses to the school administrator questionnaire), the number of schools in the school file is 1,035. The 1,035 schools are those for which both school administrator data and data from at least one student are available for the school.

7.2.1 Identification Codes

The first variable on the raw data file, STU_ID, is a unique but randomized seven-digit student identification code consisting of a five-digit school ID followed by a two-digit student code. Both sets of numbers have been randomly assigned to maintain confidentiality. Data for the four components of NELS:88 may be linked through the ID's of each component. The parent record contains the student ID. The first field of the teacher identification is the student ID. Thus, the school ID is embedded in the first five digits of each component ID (See Figure 7-1).

7.2.2 Student Questionnaire Information

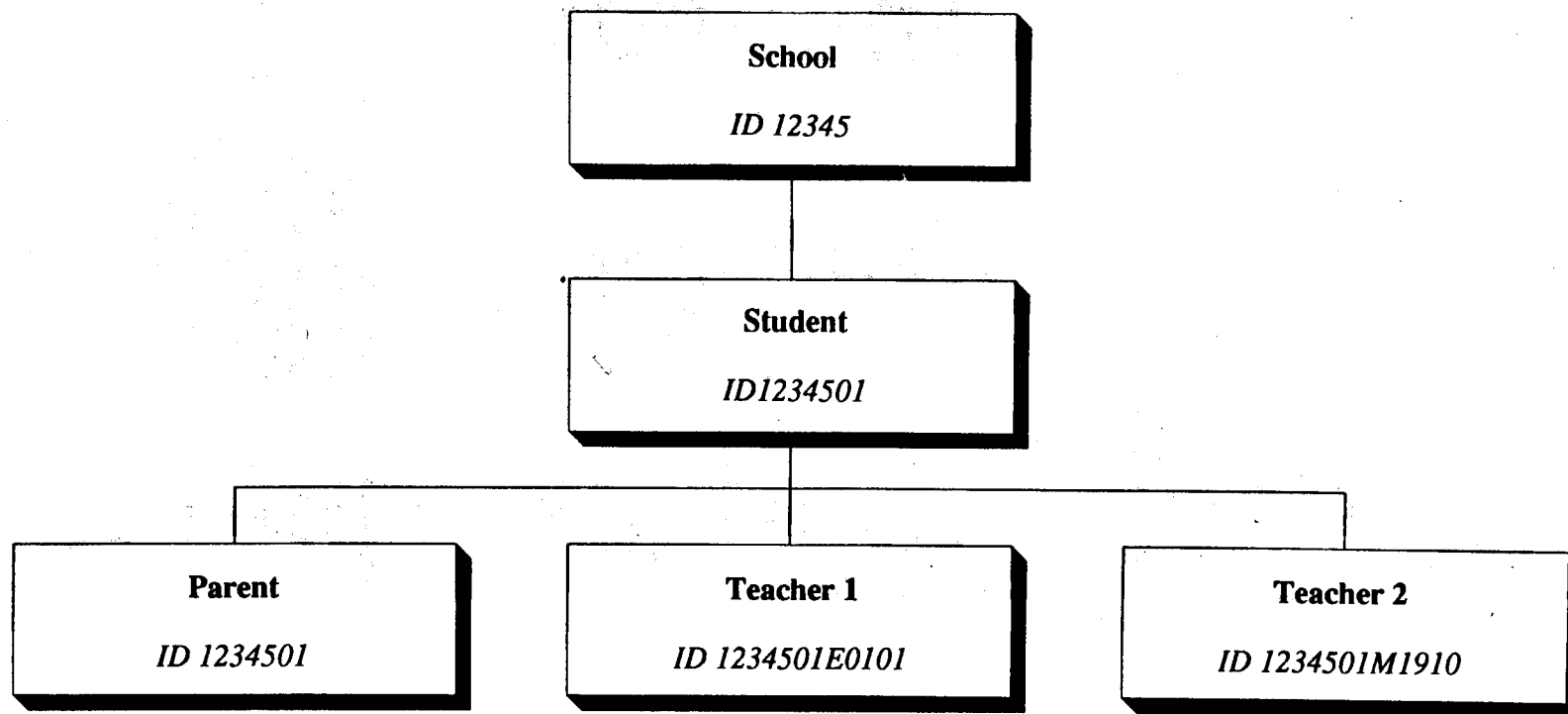
Information from the student questionnaire is presented in the same order as the questions. Variables are identified by their SAS (SPSS-X) name. All variable names begin with BYS for Base Year Student, followed by the question number. For example, BYS82K is question 82, part K, from the base year student questionnaire.

7.2.3 Sampling Weights

BYQWT is calculated from the design weight for the student (RAWWT), adjusted for the fact that some of the selected students did not complete the questionnaire. RAWWT is the reciprocal of the conditional selection probability within school for the student, given that the school was selected into the base year sample, multiplied by his or her school's design weight (SCHWT). BYQWT is included on the student data tape, as well as on the parent data tape.¹⁷ It is designed to be used in conjunction with the appropriate flag to compute population estimates of a corresponding subset of student respondents.

¹⁷ Because of the success in obtaining a parent questionnaire for such a high percentage of students, the student weight BYQWT can also be applied to provide a reasonable approximation of weighted parent statistics. See section 3.3 for details on its use.

Figure 7-1.--Data file linkages



Note: Each student was rated by teachers in two subjects. For some students, both ratings were made by the same teacher.

BYADMWT is the overall design weight for schools (SCHWT) adjusted for the fact that some of the school administrators of the participating schools did not complete a school questionnaire. BYADMWT is included on the school data tape.

7.2.4 Composite Variables

Most composite variables were constructed using responses from two or more questionnaire items. In some cases composites were constructed from numerous variables or from variables from different databases. Others were constructed by recoding a variable. A very few were simply copied from a different data source to this file for the user's convenience. All of the composite variables are described in detail in Appendix D, where they are listed along with flags and weight in the order in which they appear on the tape. Most of the composite variables can be used as classification variables or independent variables in data analysis. For this reason, composite variables may be referred to as classification variables in this or other NCES documents.

Composites of school-level characteristics provide information about the student's school.

G8TYPE classifies the type of school by the grades spanned. G8CTRL classifies the school into one of four categories, public, Catholic, other religious private, and other nonreligious private. The information for G8CTRL was taken primarily from the school data file after combining types of Catholic schools. BYSCENRL categorizes the school enrollment and G8ENROL categorizes the eighth grade enrollment as reported by the school. G8URBAN classifies urbanicity; this classification was taken directly from the QED (Quality Education Data) file, for the student's school. G8REGON indicates in which of the four U.S. Census regions the school is located. G8MINOR reflects by category the percentage of minority students in the eighth grade reported by the school. G8LUNCH reports by category the percentage of students in that student's school who receive free or reduced-price lunches. It was calculated from responses to the school questionnaire.

For some students, a school administrator questionnaire is not available. In these cases data for G8TYPE, G8CTRL, BYSCENRL, and G8ENROL were (if available) taken from the QED (Quality Education Data) file.

Some composites of school level characteristics can be considered demographic information, such as school region (G8REGON) and urbanicity of the respondent's school (G8URBAN).

Other composite and special variables. Many of the composite variables constructed were respondent demographic characteristics. SEX, RACE, HISP, API, BIRTHMO, and BIRTHYR are all examples. The SEX variable was taken first from the student questionnaire. If this source was missing or not available, then the sex variable from school rosters was used. Any records with this variable still missing had sex imputed from the respondent's name, or if that could not be done unambiguously, the value for SEX was randomly assigned. RACE also was constructed from several sources of information. The first source was the student self-report. Second, if the student information was missing or inconsistent with that of the parent, data from the parent questionnaire were used (see Appendix D). HISP (Hispanic subgroup), API (Asian and Pacific Island subgroup), BIRTHMO, and BIRTHYR were taken directly from the student questionnaire.

Socioeconomic status can be determined from BYSES and BYSESQ. The parent questionnaire was the primary source used to construct this composite, averaging the nonmissing values of five standardized components: father's and mother's educational levels, father's and mother's occupa-

tions, and family income. For cases without parent data (8.1 percent), student data were used. The first four components from the student data are the same as the components used from parent data and a ranking of material possessions was substituted for family income. BYSESQ is simply the BYSES quartile to which the respondent belongs.

Family variables include the language spoken in the home (BYHMLANG). The primary source for this composite was the student questionnaire; otherwise, parent questionnaire data were used. BYFCOMP, which categorizes the family makeup, is taken from the student questionnaire only. Additional family characteristics are available with estimated family size (BYFAMSIZ), taken first from the student questionnaire and second from the parent questionnaire, and the highest level of education reported by either of the respondent's parents (BYPARED). To construct BYPARED, student data were used whenever parent data were either missing or not available. Family variables taken from only the parent questionnaire are BYPARMAR, parent's marital status, and BYFAMINC, family income.

Four psychological scales, designed to be as comparable as possible with those on HS&B and NLS-72, were constructed from various attitude items. These scales are intended to measure locus-of-control (BYLOCUS1 and BYLOCUS2) and self-concept (BYCNCPT1 and BYCNCPT2). BYLOCUS1 and BYCNCPT1 represent only the scale items that correspond closely to NLS-72 and HS&B items. BYLOCUS2 and BYCNCPT2 represent all NELS:88 scale items. Each composite scale is the average of the standardized scores of the questionnaire items of which it is composed. For each scale a tertile ranking was calculated. These variables are named: BYLOCU1T, BYLOCU2T, BYCNCPT1T, and BYCNCPT2T. A measure of reliability, coefficient alpha¹⁸ was calculated for each of these scales. The values are: BYLOCUS1 = .5750, BYLOCUS2 = .6802, BYCNCPT1 = .7355, and BYCNCPT2 = .7867. For a list of the component items, the construction procedures, and the wording of the items in both NELS:88 and HS&B, see Appendix D. It is important to note that while the items are comparable, they are not always identical.

Educational variables include results of the cognitive tests as well as data reported on questionnaires. Eight results for each of the base year tests in the four areas of reading, mathematics, science, and social studies (history/government) are reported. The convention adopted for these thirty-two variables names is: BYTX (base year test) followed by R for reading, M for mathematics, S for science, and H for history (social science), ending with the result designator NR for number right, NW for number wrong, NNA for number not attempted, FS for formula score, STD for standardized score, IRR for IRT (Item Response Theory)-estimated number right, IRS for IRT-estimated formula score, and Q for quartile (1=low). For example, BYTXSNNA is the number not attempted on the science test. In addition, a standardized test composite for reading and math (BYTXCOMP) and its quartile (BYTXQURT) were constructed.

Seven ratings are reported that characterize the student's proficiency in reading and mathematics. These variable names begin with BYTX for base year test, followed by R for reading or M for mathematics. The variables are:

BYTXRPL1	reading proficiency level 1
BYTXRPL2	reading proficiency level 2

18 Cronbach, L. J., "Coefficient Alpha and the Internal Structure of Tests," *Psychometrika*, 16, 297-334 (1951).

BYTXRPRO	overall reading proficiency
BYTXMPL1	mathematics proficiency level 1
BYTXMPL2	mathematics proficiency level 2
BYTXMPL3	mathematics proficiency level 3
BYTXMPRO	overall mathematics proficiency

A description of the proficiency levels and an interpretation of the overall proficiency ratings are in Appendix D.

BYGRADS is an average, with all non-missing elements equally weighted, of the self-reports for grades over the four subject areas. The source is student questionnaire item 81. BYGRADSQ is the quartile distribution of BYGRADS. BYPSEPLN characterizes the postsecondary education plans of the student and was taken directly from the aspirations stated by the student in response to BYS45.

BYHOMEWK categorizes the total amount of time the student reported spending on homework a week.

BYLEP specifies whether the student has Limited English Proficiency. It was constructed from the student self-evaluations and the teacher evaluations for proficiency in using the English language. BYLM was constructed from teacher and student reports and specifies whether the student is classified as Language Minority (from a home in which a language other than English is typically spoken).

NOMSECT is the classification of the school the student expects to attend for tenth grade. The classifications were taken directly from the student data file, coded, and matched to the QED (Quality Education Data) files.

HEARIMP indicates if the student was reported to have a hearing impairment either by the parent or by the project staff as part of the survey activity. Also, the student was classified as hearing-impaired if reported as such to the Department of Education and currently mainstreamed with regular hearing eighth grade students for English or mathematics classes. This variable is less strictly defined than BYIEPFLG.

HANDPAST was constructed from responses on the parent questionnaire and indicates whether the student has ever participated in a program for the handicapped including physical, emotional, mental or learning disabilities. BYHANDPR reflects responses on the parent questionnaire and indicates whether the student is currently participating in a program for the orthopedically handicapped or learning disabled. BYHANDTR was constructed from responses on the teacher questionnaire(s) and indicates whether at least one teacher reports a handicap that interferes with school performance.

7.3 Guide to the Codebook

The codebook provides a comprehensive description of the student data file. For each variable on the tape the codebook provides a summary of the related information. The question number and wording, the variable's tape position and format, and the responses to the item along with their un-weighted frequency and percent and weighted percent are shown. See Figure 7-2 for an example. Each portion of the example is numbered. These numbers can be used to reference the associated explanation in the text following the figure.

Again, it is worth noting that there were cases when information not provided by the school administrator or the student was obtained from other sources. One example is when information from the QED datafile, used to create the sample frame, was also used to fill in missing information about the grade range of the school. Similarly, information on the student's sex and race were obtained from the school rosters when they were missing from the student questionnaires. A full description of these substitutions is in Appendix D. In addition, as noted in section 3.4, certain responses were imputed logically, as the result of machine cleaning. In general, however, there were no other attempts at imputing data for missing values. Because of this, nonresponse bias may be a problem, especially for items with high item nonresponse. These items are documented in the item nonresponse section of the sample design report.

Figure 7-2. Codebook entry

(1) Question 46

(2) Tape Pos. 159-159

(3) Format: I1

(4) BYS46 = (5) HOW SURE THAT YOU WILL GRADUATE FROM H.S

(6) How sure are you that you will graduate from high school?
(MARK ONE)

(7)

<u>RESPONSE</u>	(8) <u>CODES</u>	(9) <u>UNWGTD FREQ</u>	(10) <u>PER- CENT</u>	(11) <u>WGTD PCT</u>
Very sure I'll graduate	1	20065	81.6%	82.5%
I'll probably graduate	2	3844	15.6%	15.7%
I probably won't graduate	3	255	1.0%	1.1%
Very sure I won't graduate	4	168	.7%	.7%
RESERVED CODES:				
MULTIPLE RESPONSE	6	3	.0%	(MISS)
MISSING DATA	8	264	1.1%	(MISS)
TOTALS:		24599	100.0%	100.0%

Explanations:

(1) Question number: For variables taken directly from questionnaires, this is the question number in the original document. Composite variables and other items such as flags and weights have variable names that represent their content.

(2) Tape position: This item gives the starting and ending tape position for each variable on the data tape.

(3) Variable format: This item indicates the type of variable, its width, and the number of positions following the implicit decimal point, if any.

(4) SAS and SPSS-X variable name: Each variable on the data tape is identified by a unique SAS and SPSS-X variable name. Data indicators (such as flags and status codes) and composite variables

are given mnemonics that help identify them, for example, G8REGON for "Grade 8 in what US Census Region" and BYSES for "base year socioeconomic status." For all variables the user should be careful always to refer to the variable by its SAS (SPSS-X) variable name in any computing procedures, rather than by its question number.

(5) SAS (SPSS-X) variable label: A short variable label appears after the variable name. This label is the same as that which appears on the SAS (SPSS-X) data definition cards included on the tape.

(6) Original question wording: This reproduces the exact question wording as it appeared in the questionnaire.

(7) Response categories: This item provides either the original response categories (in the case of questionnaire items) or the recoded or constructed response categories (for composite variables and data indicators, such as flags). For display in the tables, some continuous variables have been recoded to collapse all valid values into a single response category. This allows the codebook tables to show the frequency counts, unweighted percentages, and adjusted weighted percentages for continuous variables without printing each distinct value that the variable can take. These value labels are not the same as those on the SAS (SPSS-X) data definition cards. Condensed value labels that do not cause truncation problems are provided with the data definition cards.

(8) Response codes: This item provides the actual numerical codes that appear on the data tape in the tape position specified (except for continuous variables, where the actual values that appear on the tape have been recoded to produce the frequency counts and percentages). Certain codes, discussed below, are reserved to indicate missing data, legitimate skip, and so forth.

(9) Frequency counts: This item shows the unweighted frequency counts for all records that were processed, including records that have missing data codes, legitimate skips, and so forth. Frequency counts include only those participating in the base year survey.

(10) Unweighted percentage frequencies: This column displays the frequency counts of item 9 as percentages. All records that were processed are included.

(11) Weighted percentage frequencies: This column displays percentages based on response counts weighted up to the relevant population. Cases with reserve code values are excluded from the computation.

(12) Reserved codes: In this data set certain codes, termed "reserved codes," have been chosen always to stand for certain situations. NORC and Westat have different values for reserve code 6. The student and parent surveys use NORC's convention of 6 = multiple response as shown below. The school and teacher surveys use Westat's code of 6 = don't know. Reserve codes 7, 8, and 9 are identical for all study components. These reserve codes and their interpretations are:

6 = multiple response.... more than one response where only one response was called for

7 = refusal..... respondent refused to answer an item or refused to resolve
a multiple response where only one was called for, either at the time of
the questionnaire administration or at telephone follow-up

8 = missing data..... data that should be present for this respondent is missing, but respondent did not necessarily refuse to provide data

9 = legitimate skip..... because of responses to preceding filter questions, data for this item should not be present for this respondent; that is, the value is legitimately missing

These reserved codes correspond identically to those used in NLS-72 and in the HS&B study. The codes as listed above apply to variables with single-column data fields. For variables with fields greater than one column, the left most columns are filled with 9s (e.g., 96, 996, 9996).

